Design Protocol & Notes v4

Summary

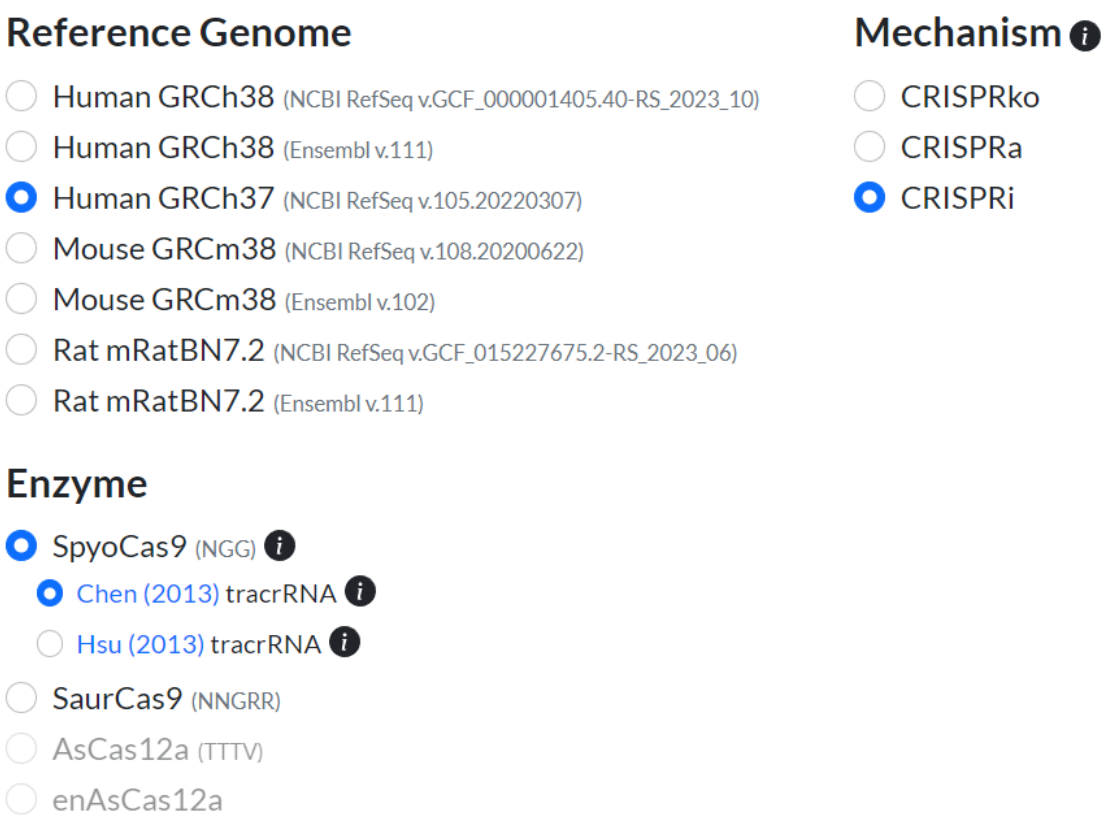
1. Preprocess SNP into regions for CRISPick input
   preprocessCRISPick.Rmd
2. Preprocess raw CRISPick output:
   preprocessGDO.Rmd
   1. Select relevant columns
   2. Calculate guide locations
3. Pick top guides:
   libraryDesignFunctions.R
   pickComparison.Rmd
   finalDesign.Rmd
   1. Final Parameters:
      1. Initial region – 200bp wide
      2. Use CRISPRi library design + Pick Quota of 30 (CRISPick parameters)
      3. On-Target Efficacy Filter > 0.2
      4. Pick guides for even strand representation (i.e. ~half of guides target each strand)
      5. Guide overlap leniency of 4bp (overlapLeniency parameter = 2)

Designed Library:

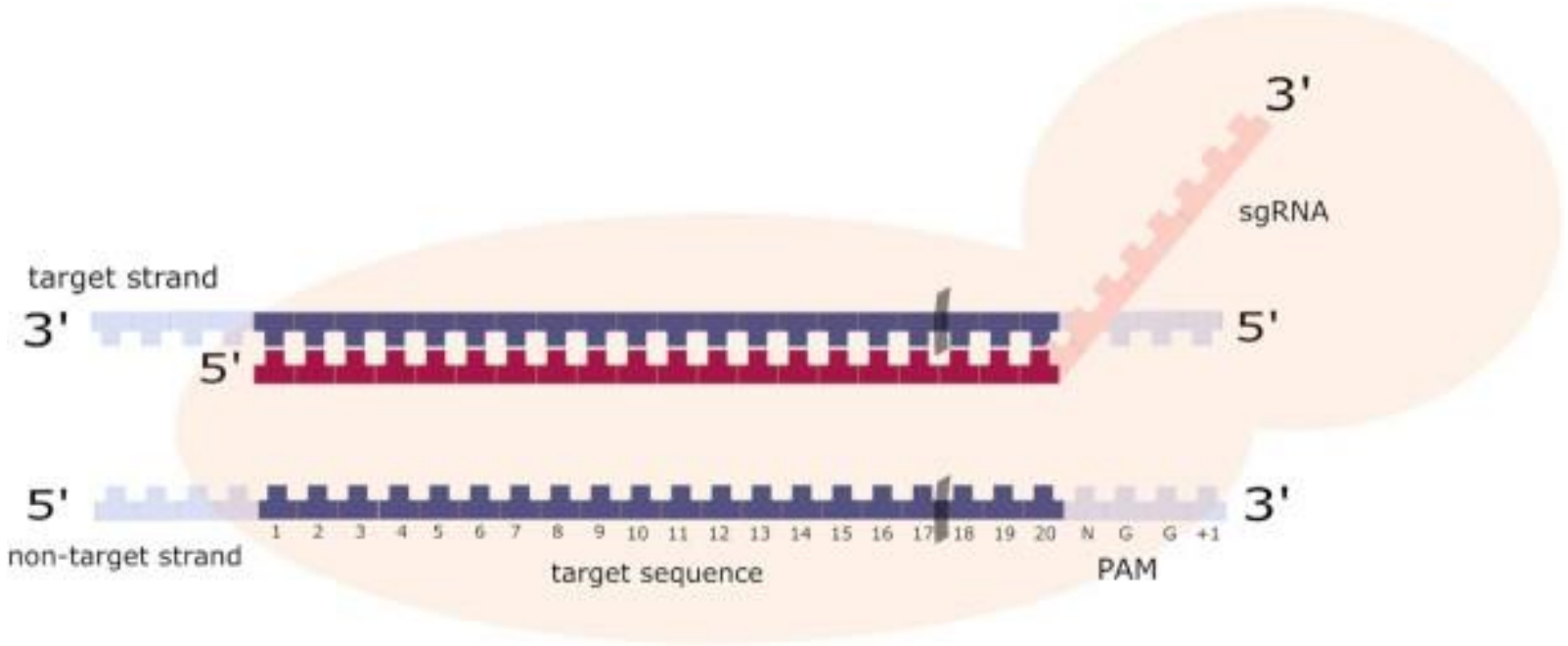
- finalTargetGDOi30.xlsx:
  contains list of targeting guides that should be good quality
- finalPromoterGDOa30.xlsx:
  contains verified list of promoter targeting guides

Preprocessing of Guides

1. Create SNP regions (+/- 100-200 around each SNP).
   Merge SNPs that are close together into the same region.
   (i.e. they have overlapping regions)
2. Use CRISPick to design guides
   for each region.
3. Filter out guides with:
   1. GC% < 0.25
   2. GC% > 0.75
   3. TTTT+ in a row
   4. On-Target Efficacy Score < 0.2



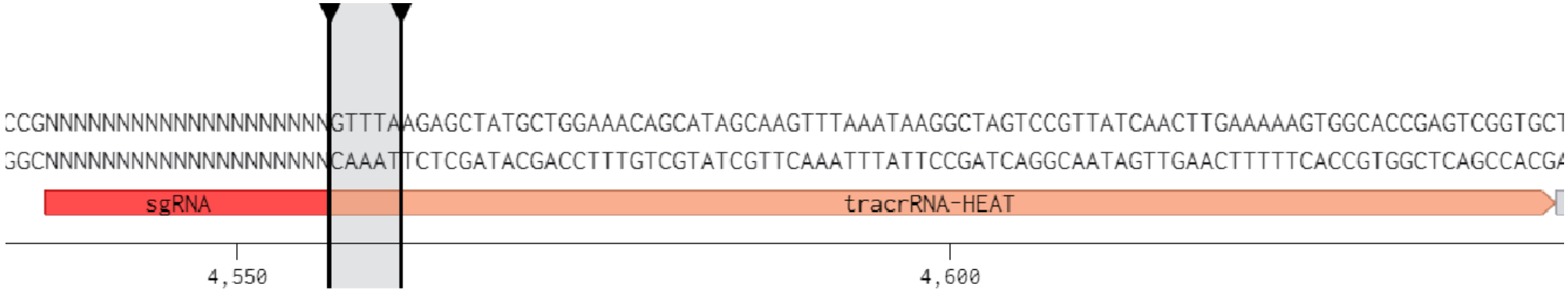
CRISPR Cas9 Cut Site



Cas9 cuts between the 17th and 18th bp of the sgRNA

[CRISPR–Cas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning - PMC] (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9023298/)

Cellecta's tracrRNA – GTTTV (Chen 2013)



Enzyme

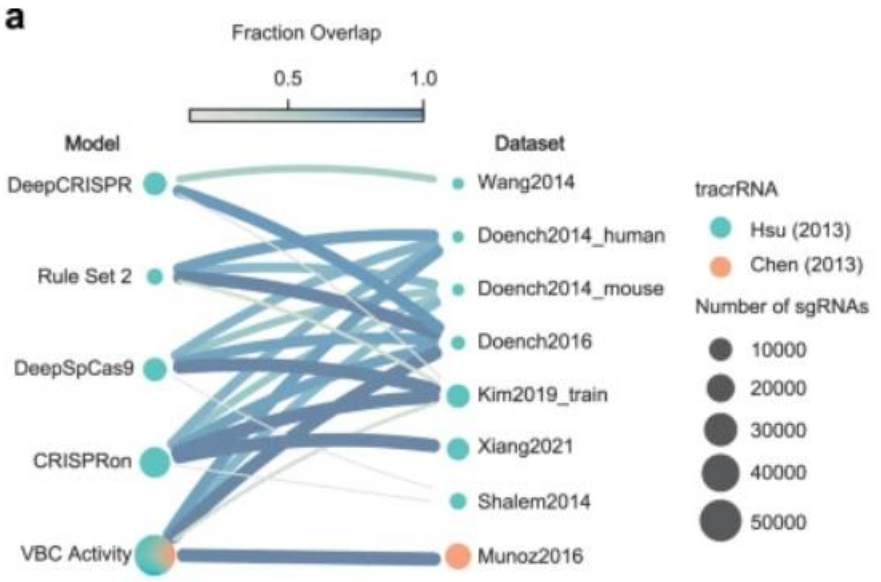
Use if the tracrRNA in your vector starts with GTTTV (V = not T). See 'How CRISPick works' for more information.

- Chen (2013) tracrRNA
- Hsu (2013) tracrRNA

Enzyme

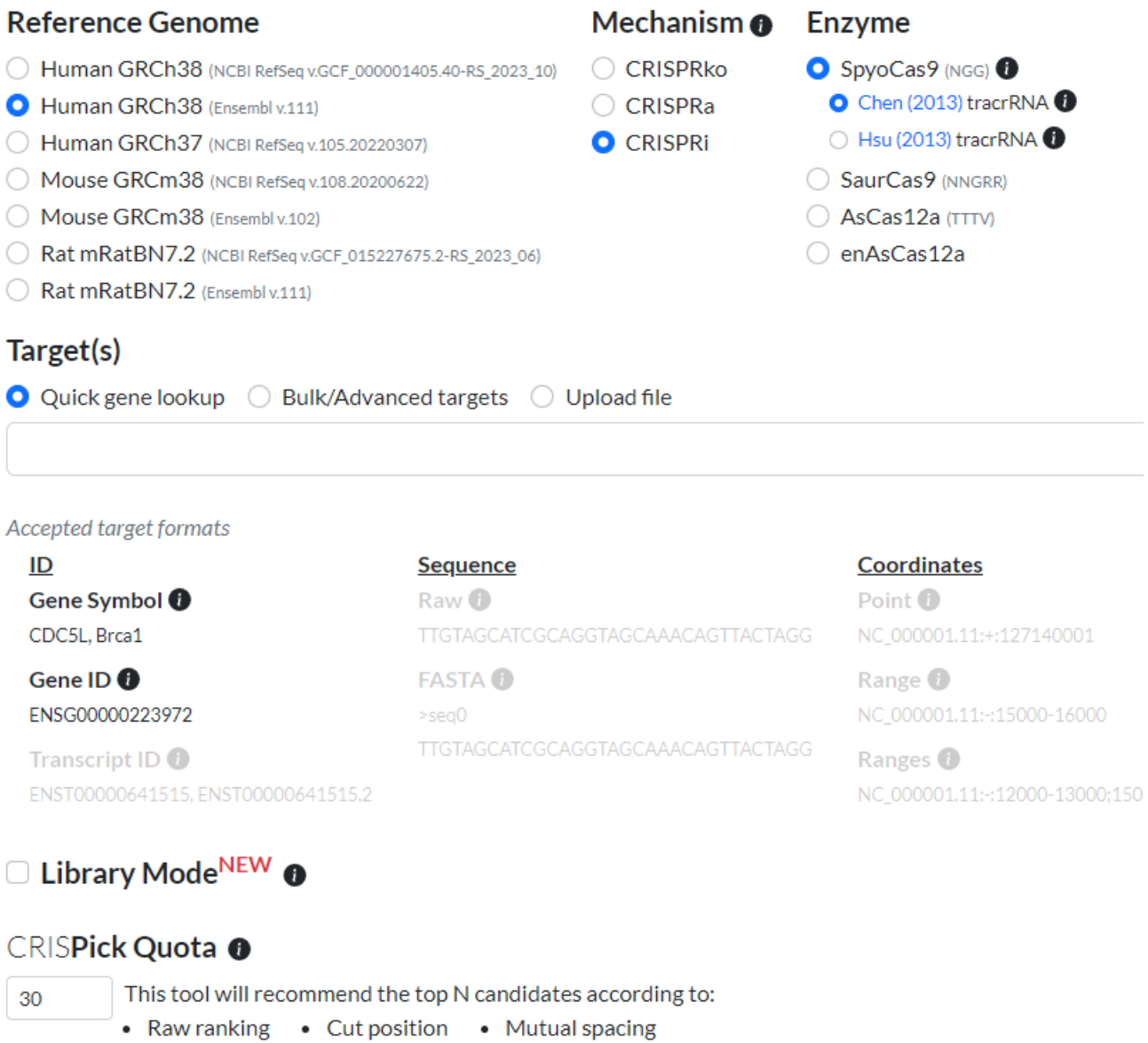
Use if the tracrRNA in your vector starts with GTTTT. See 'How CRISPick works' for more information.

- Hsu (2013) tracrRNA



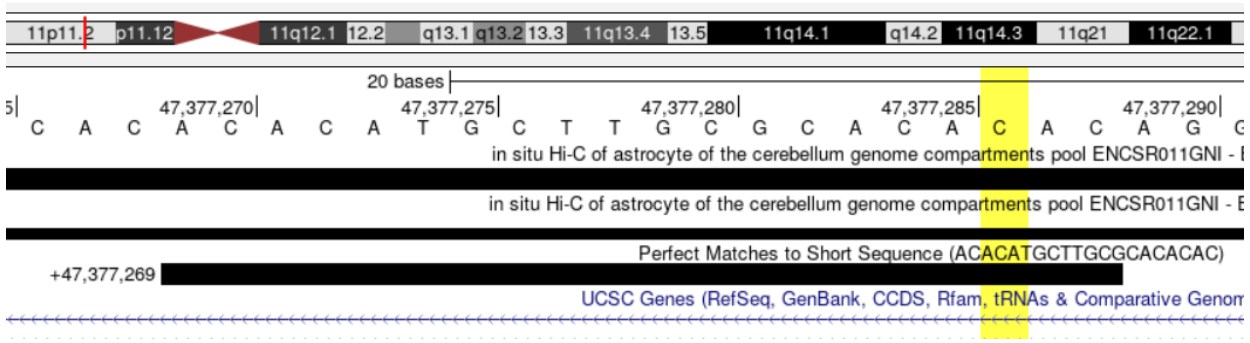
Most of the initial training datasets used the GTTTT tracrRNA.

CRISPick Settings

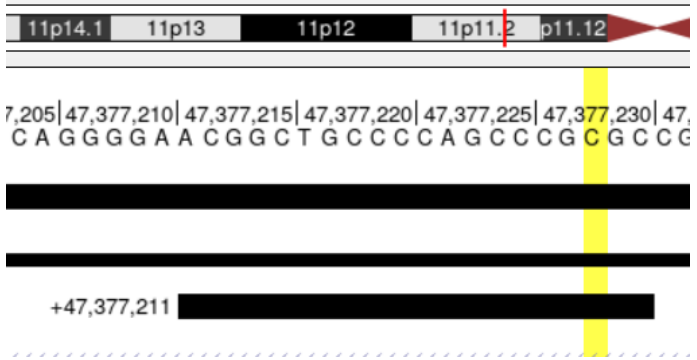


CRISPick – sgRNA locations: preprocessGDO.Rmd

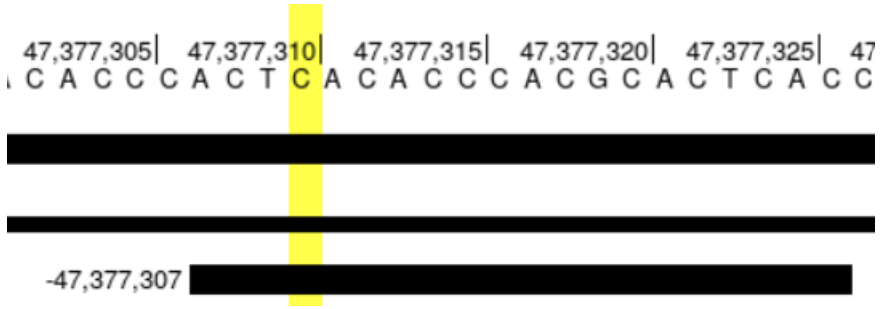
Sense Guide: 3’ ACACATGCTTGCGCACACAC 5’



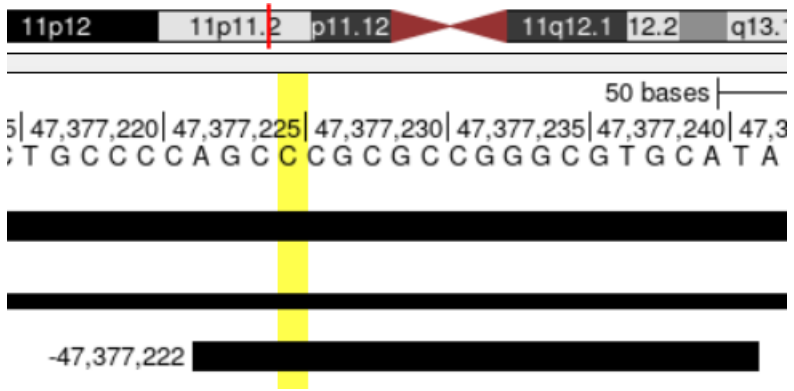
Sense Guide: 3’ ACGGCTGCCCCAGCCCGCGC 5’



Antisense Guide: 5’ GTGAGTGCGTGGGTGTGAGT 3’



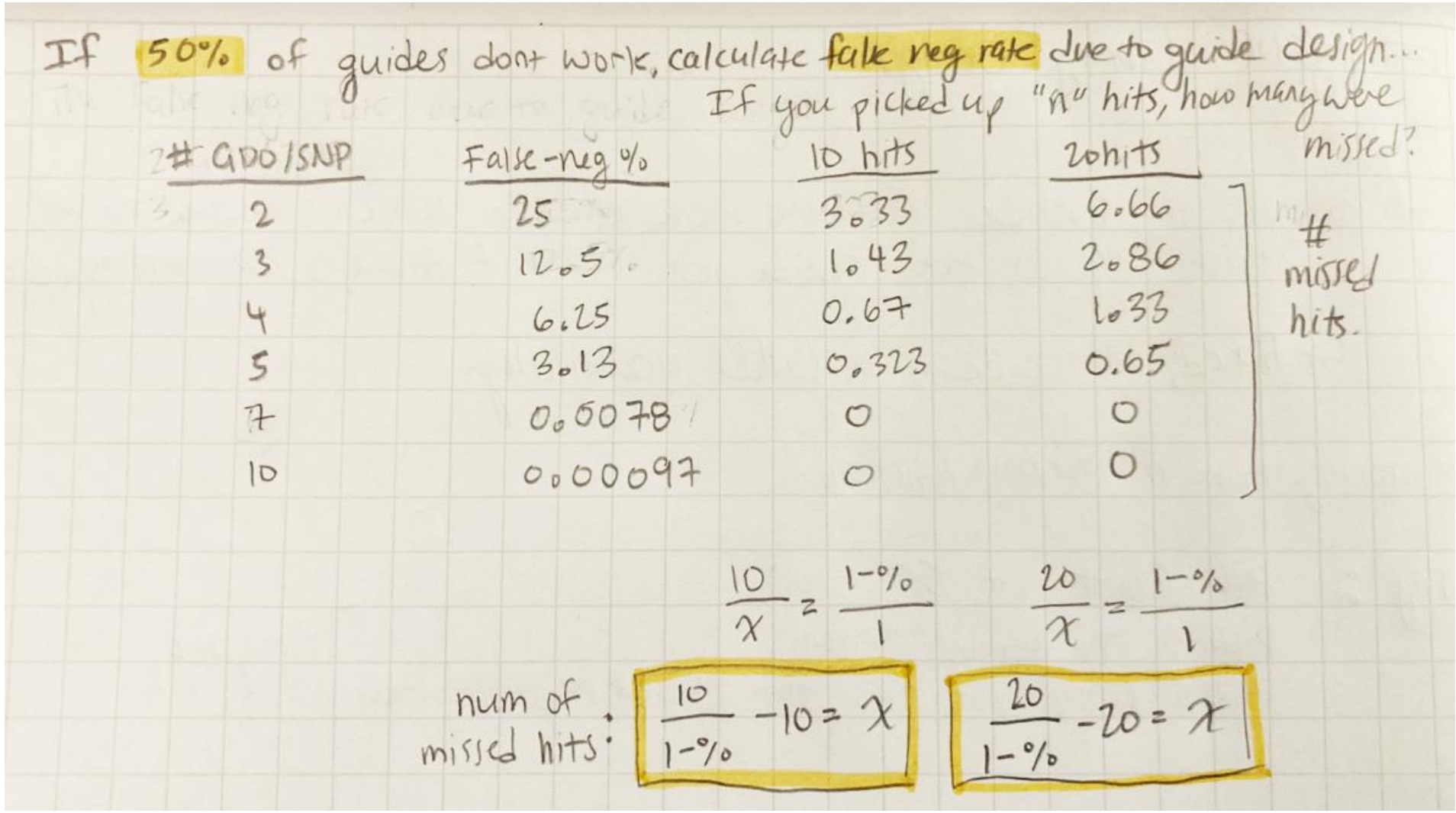
Antisense Guide: 5’ ATGCACGCCCGGCGCGGGCT 3’



Set the start and end coords for each guide depending on if they were sense or antisense.

Used GenomicRanges R package to find guides that overlap one another.

Picking number of guides per target



If 50% of guides dont work, calculate fake neg rate due to guide design...
If you picked up "n" hits, how many were missed?

# GDO/SNP	False-neg %	10 hits	20 hits
2	25	3.33	6.66
3	12.5	1.43	2.86
4	6.25	0.67	1.33
5	3.13	0.323	0.65
7	0.0078	0	0
10	0.00097	0	0

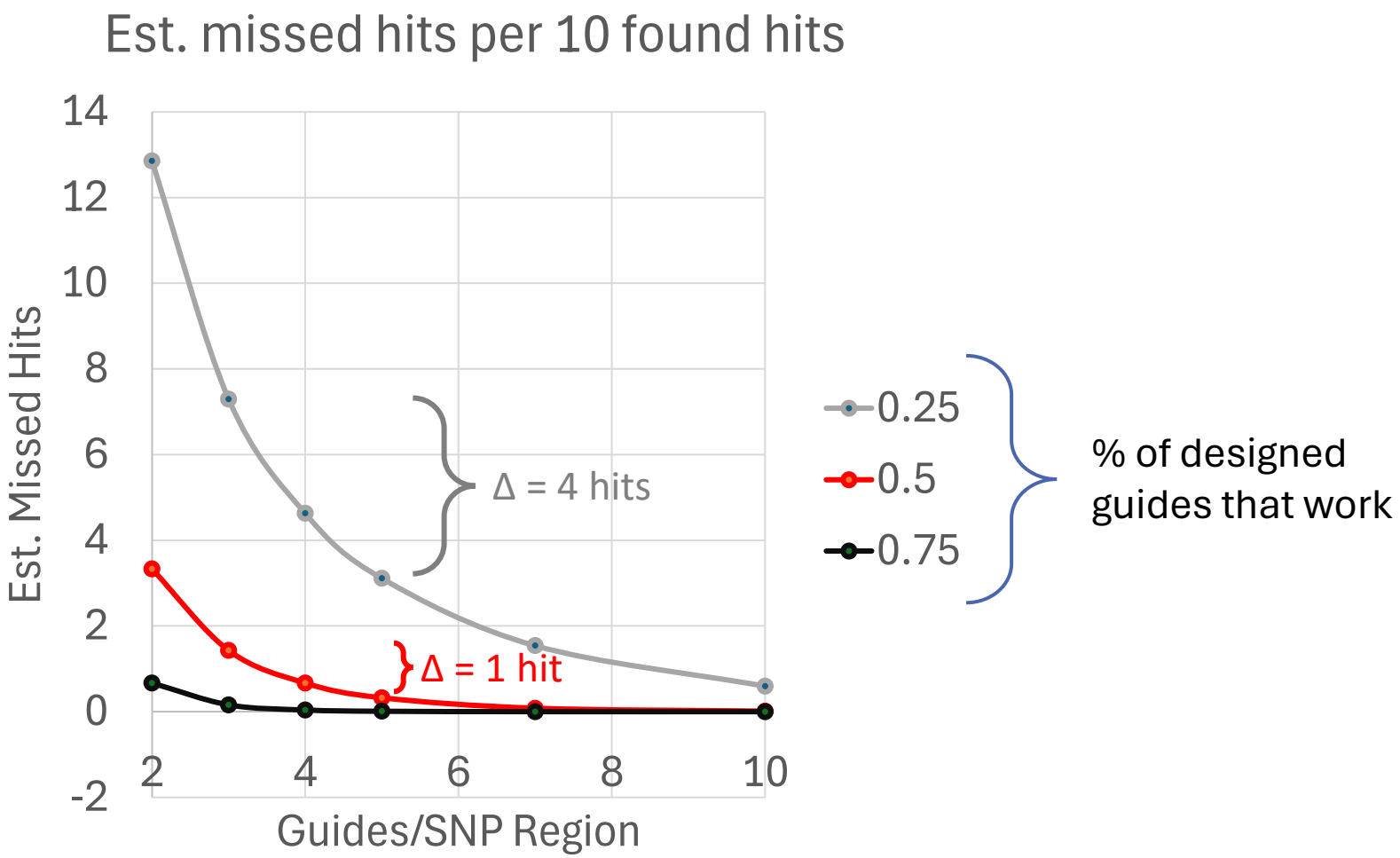
missed hits.

$$\frac{10}{x} = \frac{1-\%}{1} \qquad \frac{20}{x} = \frac{1-\%}{1}$$

num of missed hits:

$$\frac{10}{1-\%} - 10 = x \qquad \frac{20}{1-\%} - 20 = x$$

Estimating number of missed hits based on probability of designing a working guide.



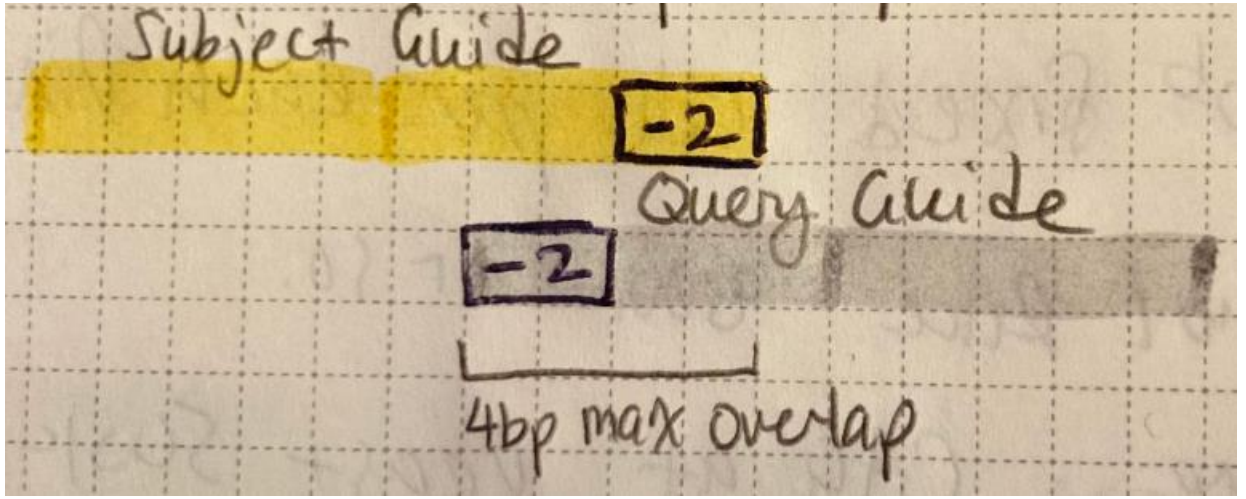
Picking Non-Overlapping Guides

For each region we have a **guides_pool** of ~30 guides/region:

1. Pick the top guide based on Pick Order (saved to **top_guides**)
2. Remove top guide + any guides that overlap the top guide from **guides_pool**. We use an overlap leniency of “2”, such that up to 4bp can overlap between any two guides – see next slide.
3. Go back to step 1
    1. Iterate until you have picked the amount of guides you want (**num_top_guides**), usually 3 or 5 or so
    2. OR, there are no more guides that have been designed for that region

Overlap Leniency Parameter

- For example, a “-2” leniency,
  where we subtract two from both ends of the guides
  would result in a max overlap of 4bp.
- For each leniency, multiply the subtraction by 2 to get the max possible overlap between guides.



Filling in Bad Targeting Guides

222 expected targeting guides

"**regions200** has the following bad regions:"
"rs10792832" "rs12539172" "rs1761452" "rs2075659" "rs2245466"
"rs58250526" "rs58804619" "rs6064392" "rs672399" "rs7149638"
"rs73208737" "rs78810900" "rs9271608"

"**regions300** has the following bad regions:"
"rs10792832" "rs1761452" "rs58250526" "rs58804619" "rs7149638"

"**regions400** has the following bad regions:"
"rs10792832"

Cut off for On-Target Efficacy (5k cis)

Based on CRISPick GDO filtering guidelines, will remove any guides with **<0.2 On-Target Efficacy Score**.

CRISPRa vs CRISPRi guides from CRISPick

~30% of top guides are picked differently between a CRISPRa vs CRISPRi library design.

Between the two parameters, Off-Target Rank is different. Tier I and Tier II matches are different.

Example of the same guide but under different library design parameters:

CRISPRi: top guide is Pick Order 1

#.Off-Target.Tier.I.Match.Bin.II.Matches	#.Off-Target.Tier.II.Match.Bin.II.Matches	#.Off-Target.Tier.III.Match.Bin.II.Matches
0	0	98
1	0	67

CRISPRa: top guide is Pick Order 2

#.Off-Target.Tier.I.Match.Bin.II.Matches	#.Off-Target.Tier.II.Match.Bin.II.Matches	#.Off-Target.Tier.III.Match.Bin.II.Matches
0	1	97
0	0	68

For now, will stick with CRISPRi for design (might be a wider TSS site). Could revisit later...
Might be possible to re-engineer the Rank + Pick Order.. Maybe for future library wishlist.

Here we get the sum of all off-target matches.
Whichever library design has a larger sum, we pick that one cause it picked up more.

Comparing sum of off-target matches:

CRISPRi picks up more off-target hits.

Tier I and Tier II off-target hits are based regions relative to the TSS site.
CRISPick gets TSS locations from MANE_Select TSS annotations from NCBI.

200bp, pick 30

Lib	TI BI	TII BI	TIII BI	TI BII	TII BII	TIII BII
CRISPRi	5	4	966	2541	2010	418627
CRISPRa	2	3	970	2184	1595	416092

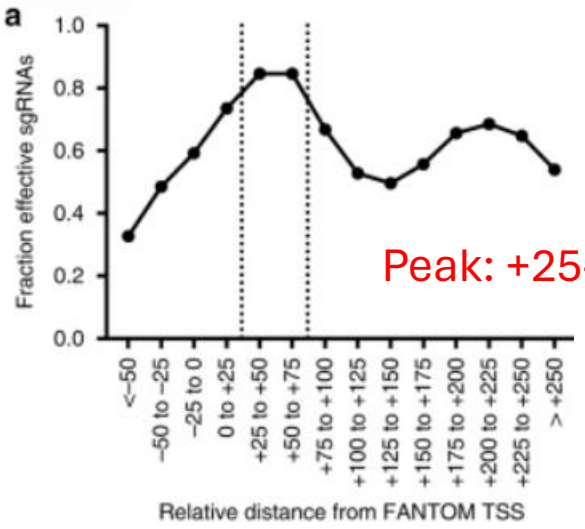
200bp, pick 50

Lib	TI BI	TII BI	TIII BI	TI BII	TII BII	TIII BII
CRISPRi	7	4	1220	4042	2757	554220
CRISPRa	3	3	1224	3466	2224	555591

Notes on optimal distance:

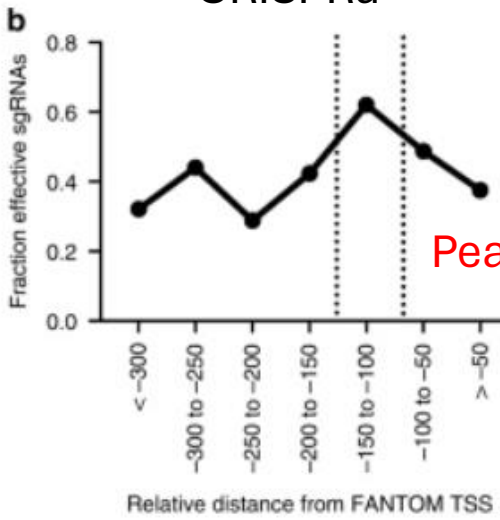
- Have ~50bp window for optimal targeting Okay up to 100bp or so.
- However, since we are targeting indirectly, harder to say what this distance should be.
- Could look at our data.

CRISPRi



Peak: +25-75bp

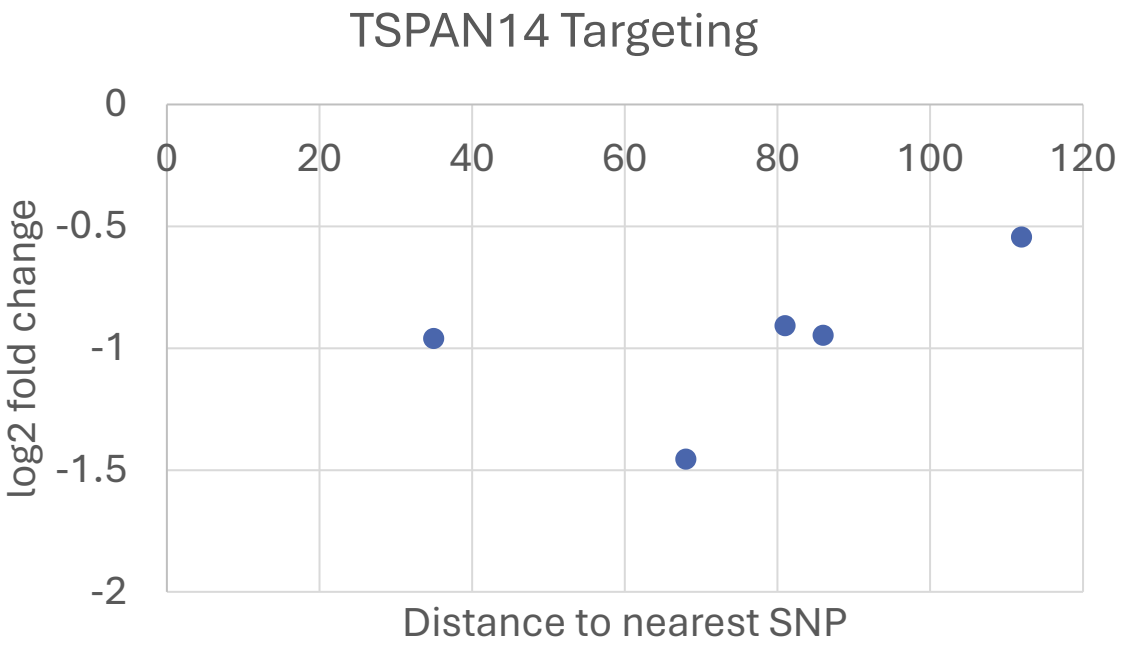
CRISPRa

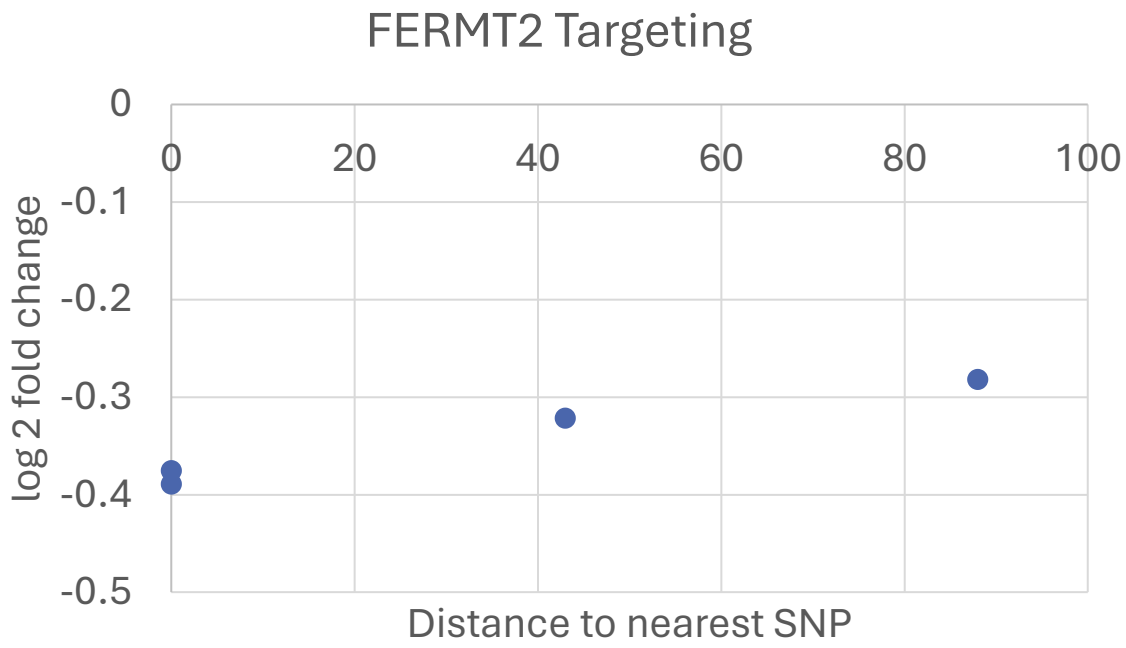


Peak: -100-150bp

Source (CRISPick 2018 paper):
[Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities | Nature Communications](https://www.nature.com/articles/s41467-018-07901-8)

Log2 fold change vs distance of guide to nearest SNP.





Comparing initial region (pick 3 guides)

Initial region: 200bp wide

[1] "regions200 has the following bad regions:"

[1] "rs10792832" "rs12539172" "rs1761452" "rs2075659" "rs2245466" "rs2279590" "rs58250526" "rs58804619" "rs6064392" "rs672399"

[11] "rs7149638" "rs73208737" "rs78810900" "rs9271608"

[1] "regions300 has the following bad regions:"

[1] "rs10792832" "rs1761452" "rs58250526" "rs58804619" "rs7149638"

[1] "regions400 has the following bad regions:"

[1] "rs10792832“

[1] "Mean Num GDO per Region: 3"

[1] "Mean On-Target: 0.57612027027027"

[1] "Reg w OffTarget, Tier I Bin I: 1"

[1] "Reg w OffTarget, Tier II Bin I: 0"

[1] "Reg w OffTarget, Tier III Bin I: 6"

[1] "Reg w OffTarget, Tier I Bin II: 35"

[1] "Reg w OffTarget, Tier II Bin II: 67"

[1] "Reg w OffTarget, Tier III Bin II: 222"

[1] **52.63514**

Initial region: 300bp wide

[1] "regions300 has the following bad regions:"

[1] "rs10792832" "rs1761452" "rs58250526" "rs58804619" "rs7149638"

[1] "regions400 has the following bad regions:"

[1] "rs10792832“

[1] "Mean Num GDO per Region: 3"

[1] "Mean On-Target: 0.61659014084507"

[1] "Reg w OffTarget, Tier I Bin I: 0"

[1] "Reg w OffTarget, Tier II Bin I: 0"

[1] "Reg w OffTarget, Tier III Bin I: 5"

[1] "Reg w OffTarget, Tier I Bin II: 23"

[1] "Reg w OffTarget, Tier II Bin II: 47"

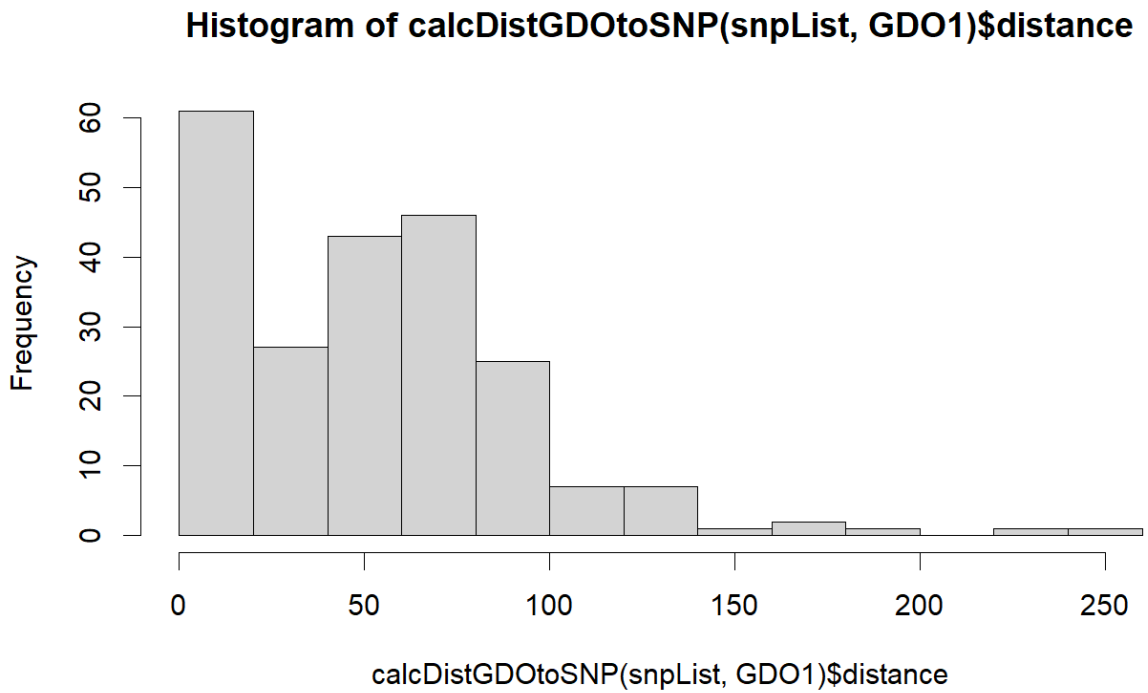
[1] "Reg w OffTarget, Tier III Bin II: 213"

[1] 72.69014

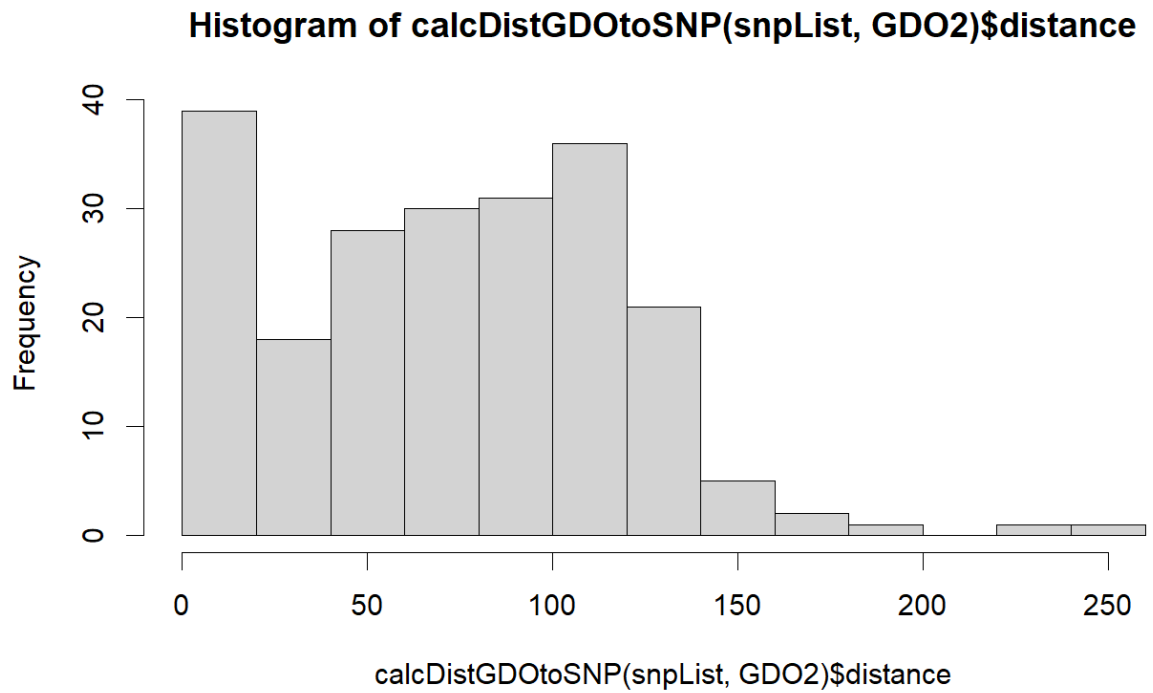
Pick an initial region of 200bp since guides are ~20bp closer to SNPs, on average.

Comparing Initial Regions

Initial region: 200bp wide



Initial region: 300bp wide



Additional Library Parameters

30 vs 50 CRISPick picking quota

- Libraries are identical

Overlap Leniency

- Slight improvement for increased overlap leniency.
- Chose a 4bp overlap leniency (overlapLeniency = 2)
- Only ~20 guides are affected.

List of Positive Control Targets

- **FERMT2 (new)**
- BIN1
- SNX1
- TSPAN14
- CLU
- RPA1/SMYD4
- **ACP2 (new)**
- 21 to 35(?) total guides; might be able to fit more if desired.

Add SEZ6L2?
More of a neuronal gene, but maybe relevant if we reuse library in neurons.

Dropping RAB1A since it's not interesting and we have some positive controls that should work and are more interesting.

(+) Ctrl Guides, guidelines (see appendix for details)

- Pick everything for CRISPRa rn.
  Guides that bind to non-template strand are okay too.
- I think CRISPRa guides are more likely to work for CRISPRi rather than vice versa.
  - And we are definitely doing a CRISPRa screen.
    CRISPRi is not sure yet.
- Can discuss and revise later, too.

- See final designs in **finalPromoterGDOa30.xlsx**.

Pilot Cost: Email from Charlly

The "single-end" read capacities are as follows for the following ILMN NovaSeq flow cells:

- S1: 1.3-1.6B
- S2: 3.3-4.1B
- S4: 8-10B

Assuming 25K reads/cell, we estimate that the following reads are needed for the following cell numbers:

- 100K cells: 2.5B reads
- 250K cells: 6.25B reads
- 500K cells: 12.5B reads

Here are pricing for the following flow cells using 200 cycle kit:

- S1: $6,925.00
- S2: $12,030.00
- S4: $16,860.00

We can discuss a number of types of strategies for sequencing, but one option is to start the first assay with 100K cells and sequence first on one S1 flow cell, which would reach ~50% of the target sequencing depth. This should provide sufficient info about the data quality and enable some prelim analyses, and if things look good, you can then do more sequencing (on another S1 flow cell). This step-wise approach would probably be more prudent than trying to sequence on an S2 flow cell upfront.

Cheers,

Charlly

Pilot Experiment Scenarios:

What is the total budget? How many exp to run?

Run 1 Experiment (100k cells)

Upfront: Library + Scale Base + S1

$12.5+$8.5+$7k = **$28k**

Looks good?

Add S1 for more seq ($7k)

$34.5k total for 1 Exp

3 Exp afterwards: +$31.4k

$65.9k total for 4 Exp; $16.4/exp.

Run 3 Experiments (375k cells)

Upfront: Library + Scale Base + S1

$12.5+$8.5+$7k = **$28k**

Looks good?

Add 2 Scale Ext + S4 ($22.8k)

$50.5k total for 3 Exp

$16.8k/exp

S1 should be enough to check half of 125k cells from base kit?

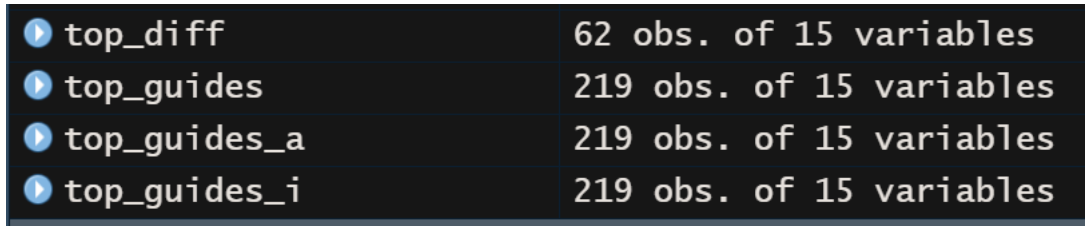
Check this with Charlly.

Appendix / Of Note

CRISPRi vs CRISPRa to search for guides..

- For:
  top_diff <- anti_join(top_guides_i, top_guides_a, by="sgRNA.Sequence")
  - 62 guides are different between the two..

- Will use CRISPRi for now so you can compare to pilot.



Checking Region Widths

- Problem? This is regions_300
- Check that you're pulling from the region width you think you are.

27	17	1639570	1640183	614	*	rs1317708|rs4989024|rs874424
28	17	1640380	1641184	805	*	rs2287322|rs8077638|rs34962442|rs620900

List of "poor regions" from 200bp width region.

Some regions get really wide when you combine nearby SNPs. These usually are easy to design guides for, though.

```
> max(regions_200$width)
[1] 463
> max(regions_300$width)
[1] 805
> max(regions_400$width)
[1] 905
```

	SNP
1	rs1761452
2	rs2075659
3	rs2245466
4	rs2452758
5	rs58250526
6	rs672399
7	rs7149638
8	rs73208737

+300bp width region

	SNP
1	rs1761452
2	rs2245466
3	rs58250526
4	rs7149638

Checking Region Widths

Poor_reg from regions_200:
Subset(regions_300, SNP %in% poor_reg$SNP)

	chr	start	end	width	strand	SNP
8	11	60019963	60020261	299	*	rs672399
31	17	18044443	18044741	299	*	rs2075659
40	19	54815217	54815515	299	*	rs1761452
44	12	113591288	113591586	299	*	rs58250526
45	12	113634906	113635204	299	*	rs73208737
62	14	53346782	53347080	299	*	rs7149638
66	5	86181554	86181852	299	*	rs2452758
71	4	40198697	40198995	299	*	rs2245466

Poor_reg after adding regions_300:
Subset(regions_400, SNP %in% poor_reg$SNP)

	chr	start	end	width	strand	SNP
40	19	54815167	54815565	399	*	rs1761452
44	12	113591238	113591636	399	*	rs58250526
62	14	53346732	53347130	399	*	rs7149638
71	4	40198647	40199045	399	*	rs2245466

Only 3 Regions in width 300 or 400 but not 200

200bp width

	chr	start	end	width	strand	SNP
27	17	1639620	1639818	199	*	rs4989024
28	17	1639866	1640133	268	*	rs1317708|rs874424
29	17	1640430	1640892	463	*	rs8077638|rs34962442|rs62090051
30	17	1640936	1641134	199	*	rs2287322

300 or 400bp width

	chr	start	end	width	strand	SNP
1	17	1639570	1640183	614	*	rs1317708|rs4989024|rs874424
2	17	1640380	1641184	805	*	rs2287322|rs8077638|rs34962442|rs62090051
3	10	82269312	82269997	686	*	rs1870137|rs1870138|rs7080009

Adding Positive Controls (MAST analysis)

Working Guides	Guides with No Effect
RAB1A-3	RAB1A-2 (close)
BIN1-2	RPA1-SMYD4-3 (close)
SNX1-2	BIN1-1
BIN1-3	CLU-1
RAB1A-1	CLU-2
TSPAN14-3	CLU-3
TSPAN14-1	RPA1-SMYD4-1
SNX1-3	RPA1-SMYD4-2
	SYVN1 (why?)
	SNX1-1
	TSPAN14-2

In /reference folder, see power_check_result for SCEPTRE analysis of control guides.

Other Todo?
Deep Learning enabled scoring algs

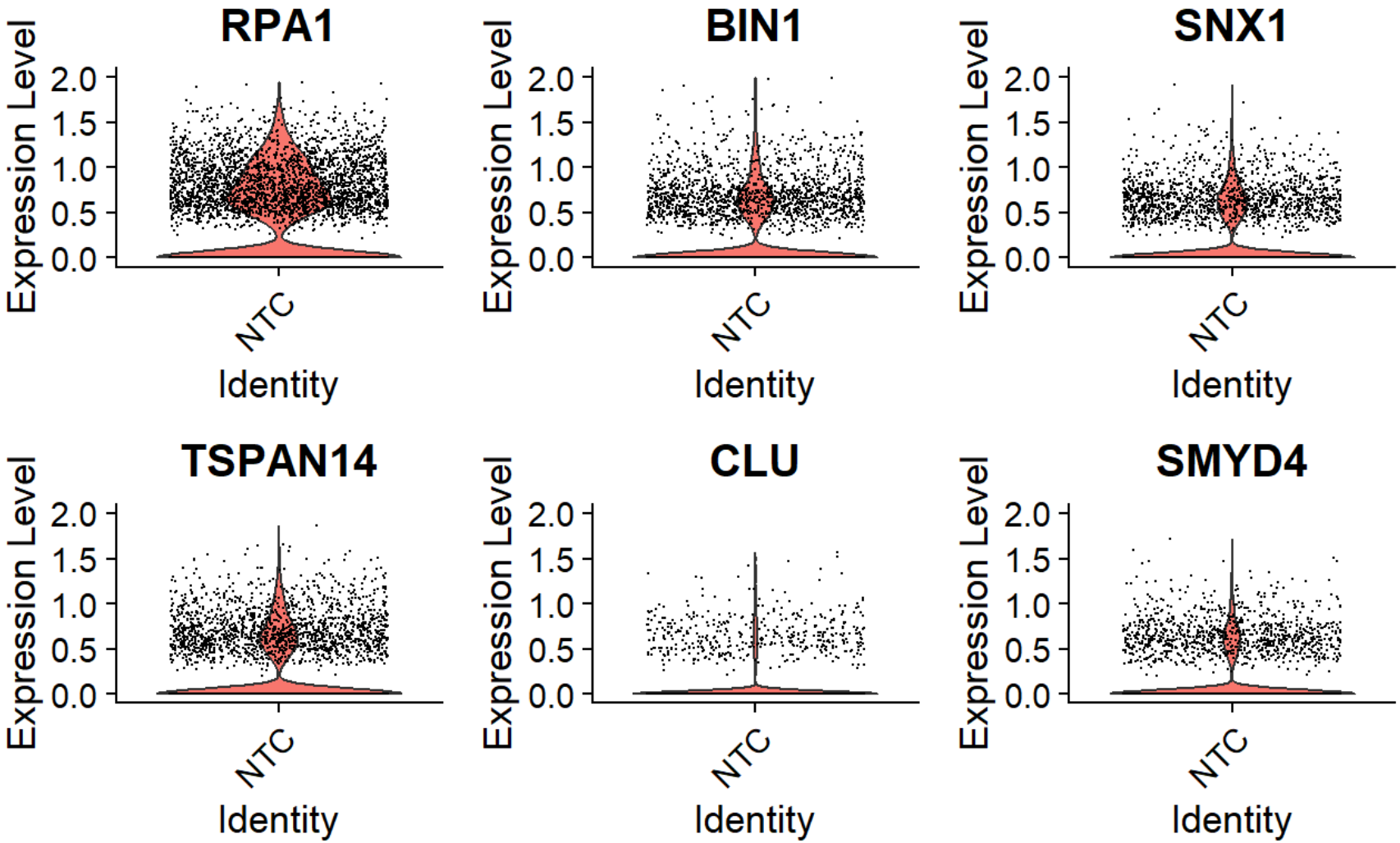
Table 6.

Predicted efficiency of each gRNA from the De Ravin study

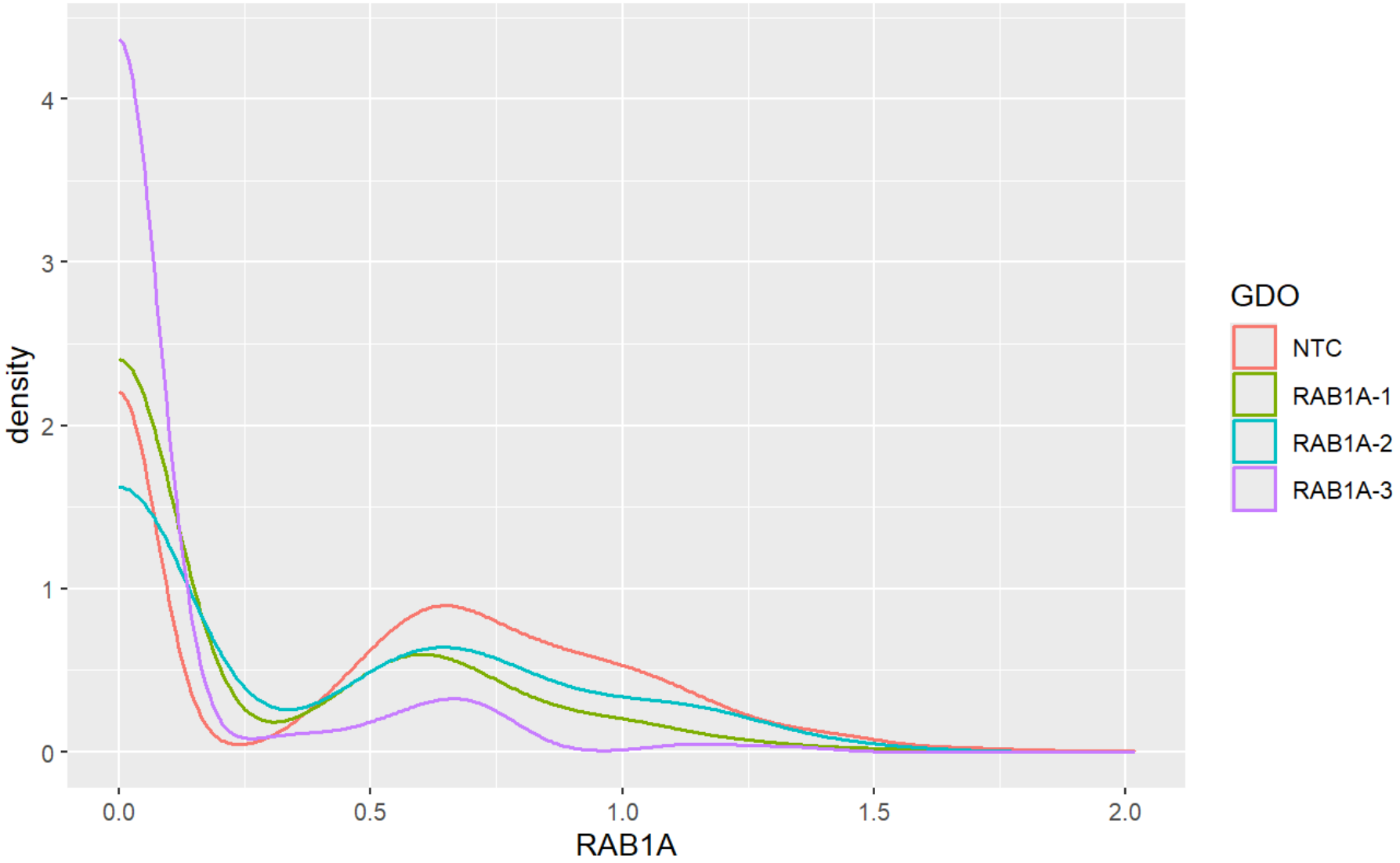
	gRNA2	gRNA3	gRNA8	gRNA1
Actual cleavage	**0.21**	0.0125	0.01	0.004
DeepCRISPR	0.099	0.431	0.096	**0.464**
DeepSpCas9	**0.629**	0.482	0.122	0.063
DeepHF	**0.628**	0.421	0.312	0.363
Average	**0.594**	0.437	0.190	0.220
CRISPRLearner	**0.559**	0.393	0.255	0.371
C-RNNCrispr	**0.202**	0.164	0.189	0.157
E-CRISP	**0.530**	0.353	0.270	0.281

[CRISPR–Cas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning - PMC](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9023298/)

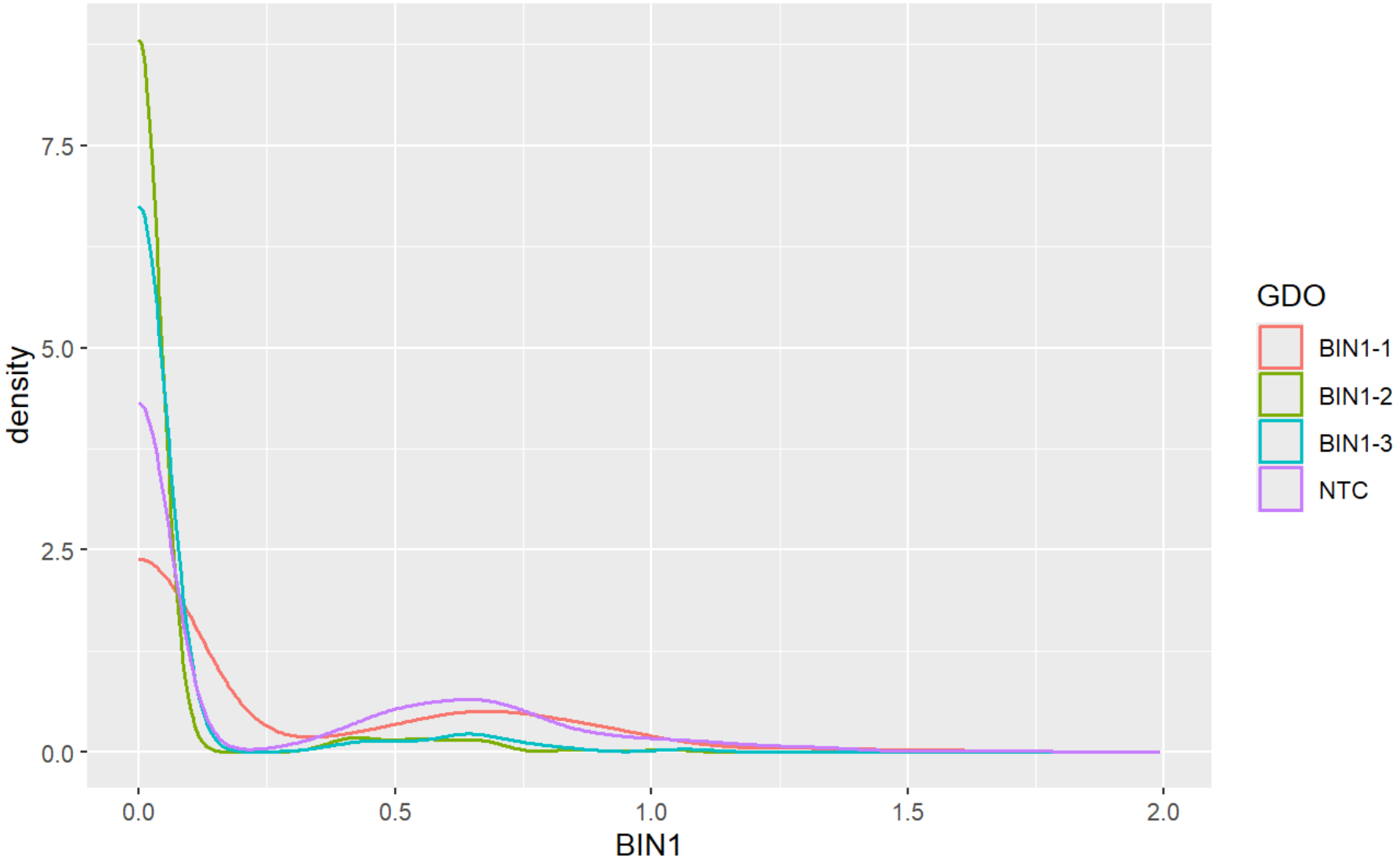
NTC Expression for each CTRL gene



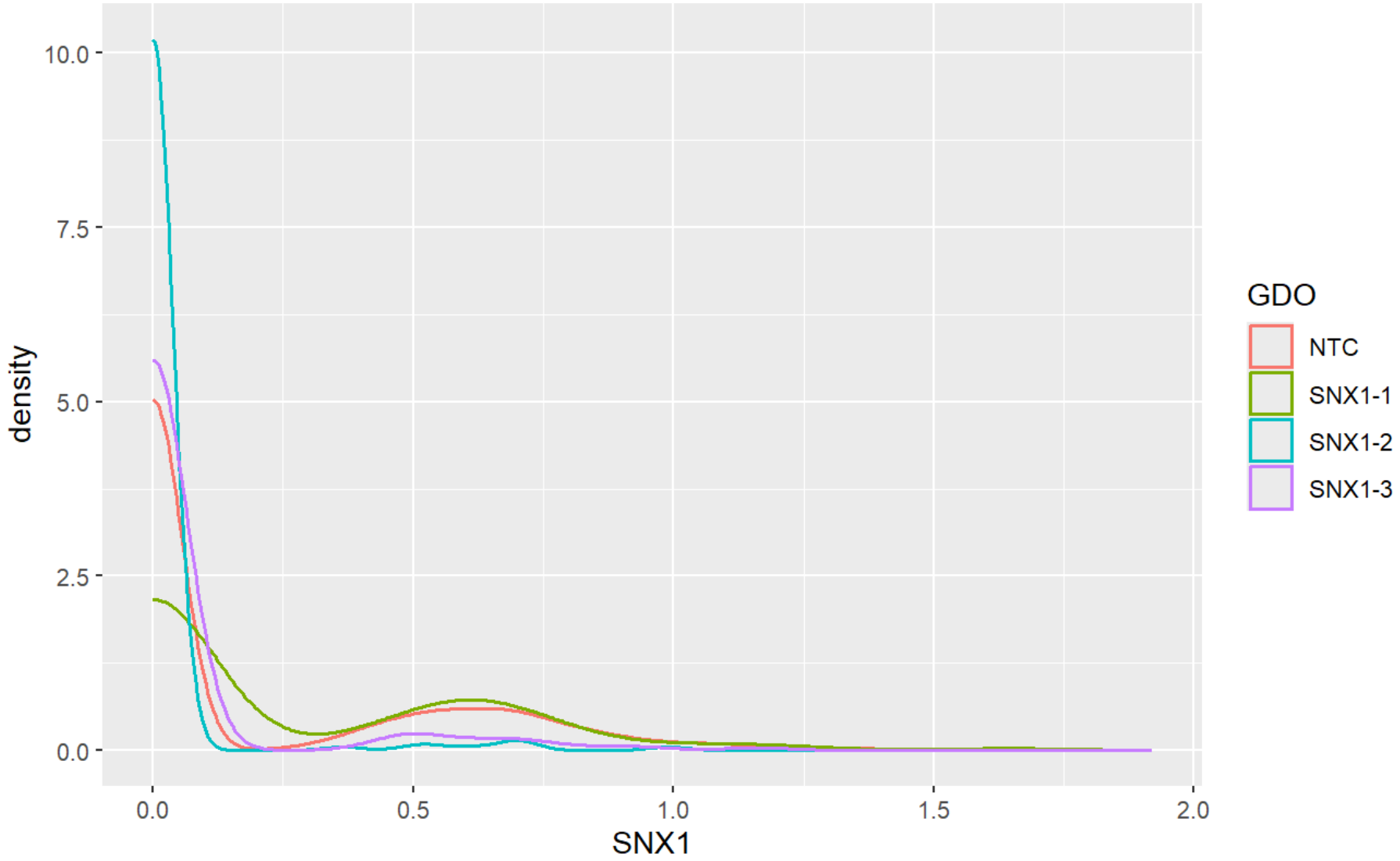
Positive Controls: RAB1A



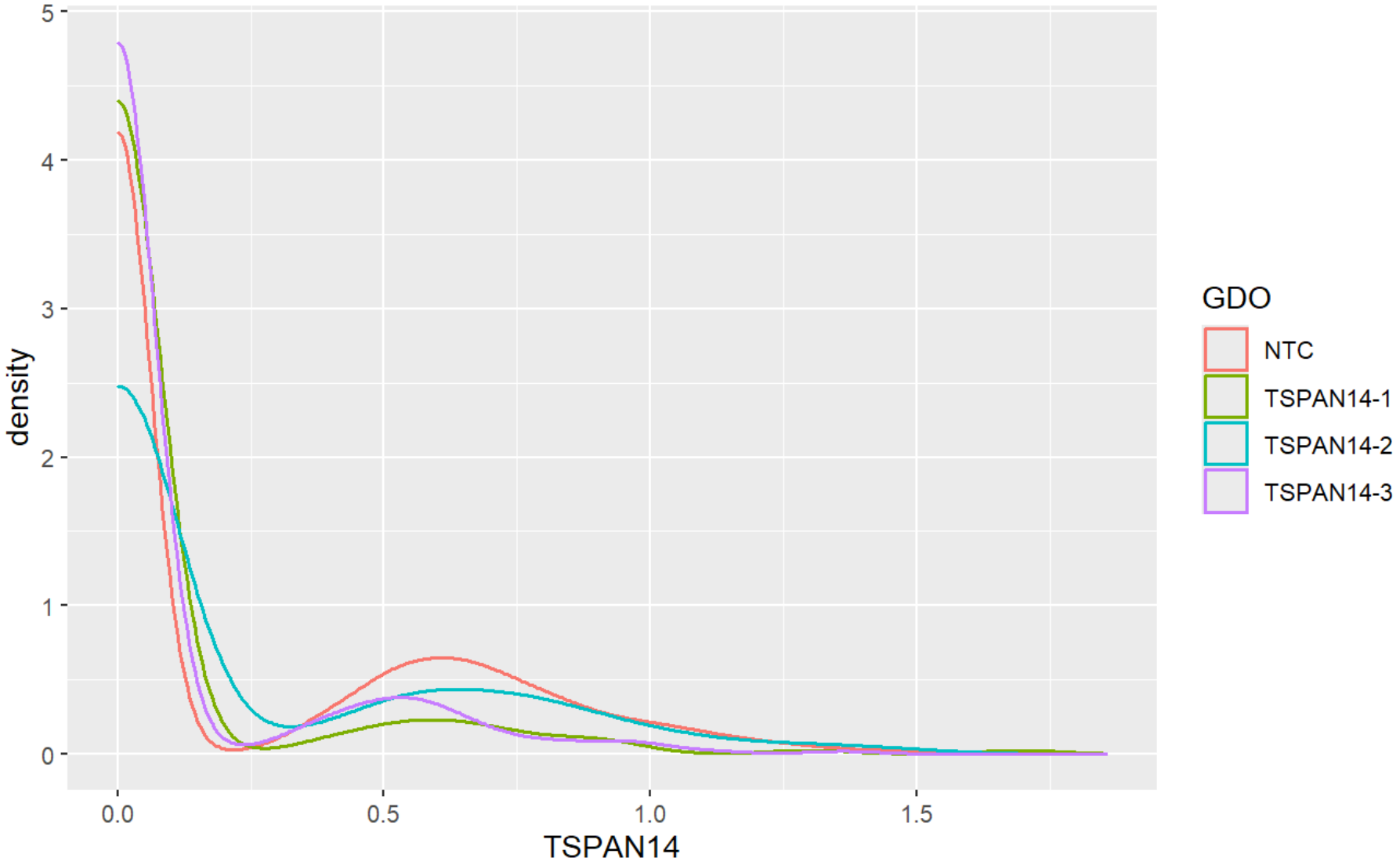
Positive Controls: BIN1



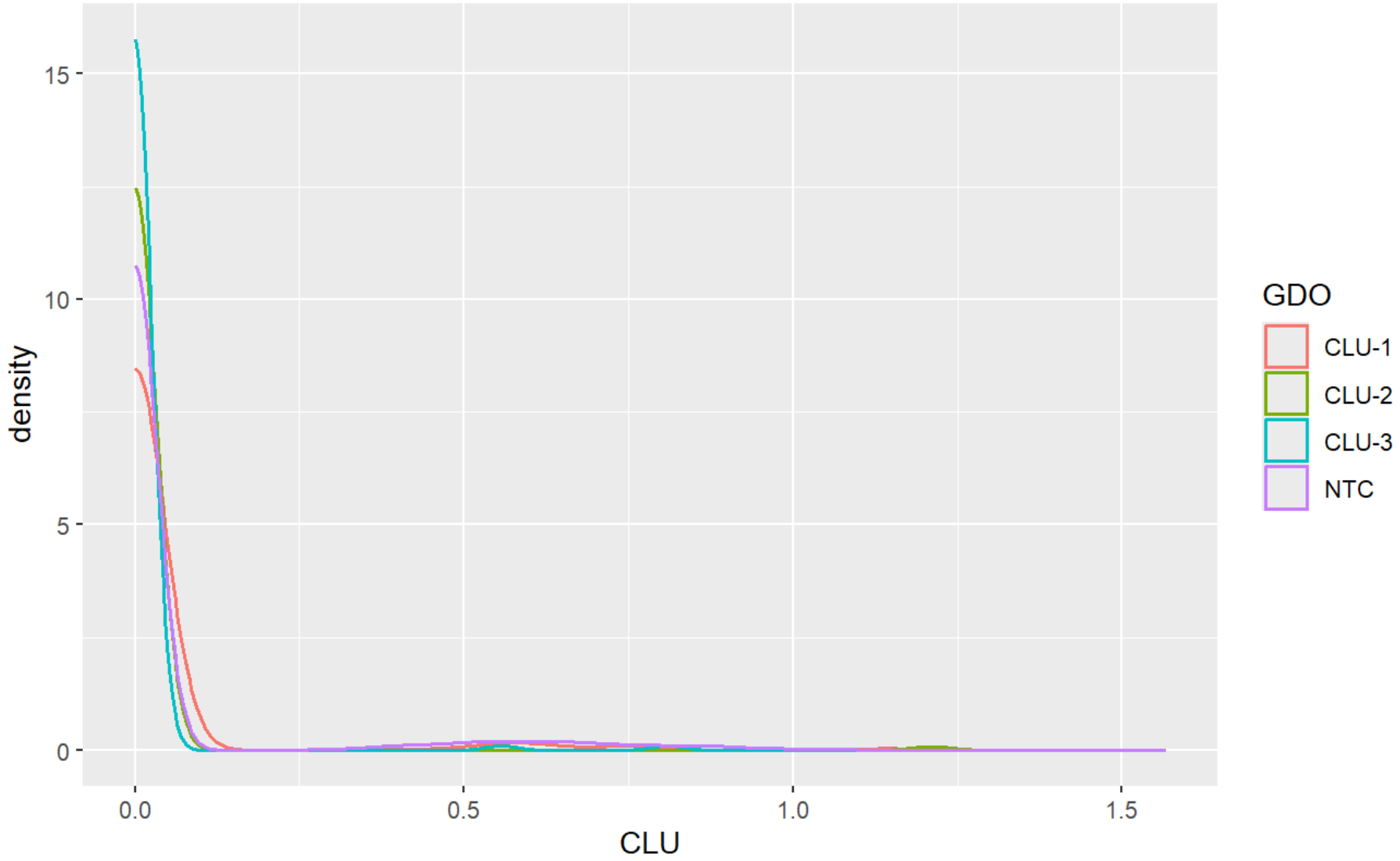
Positive Controls: SNX1



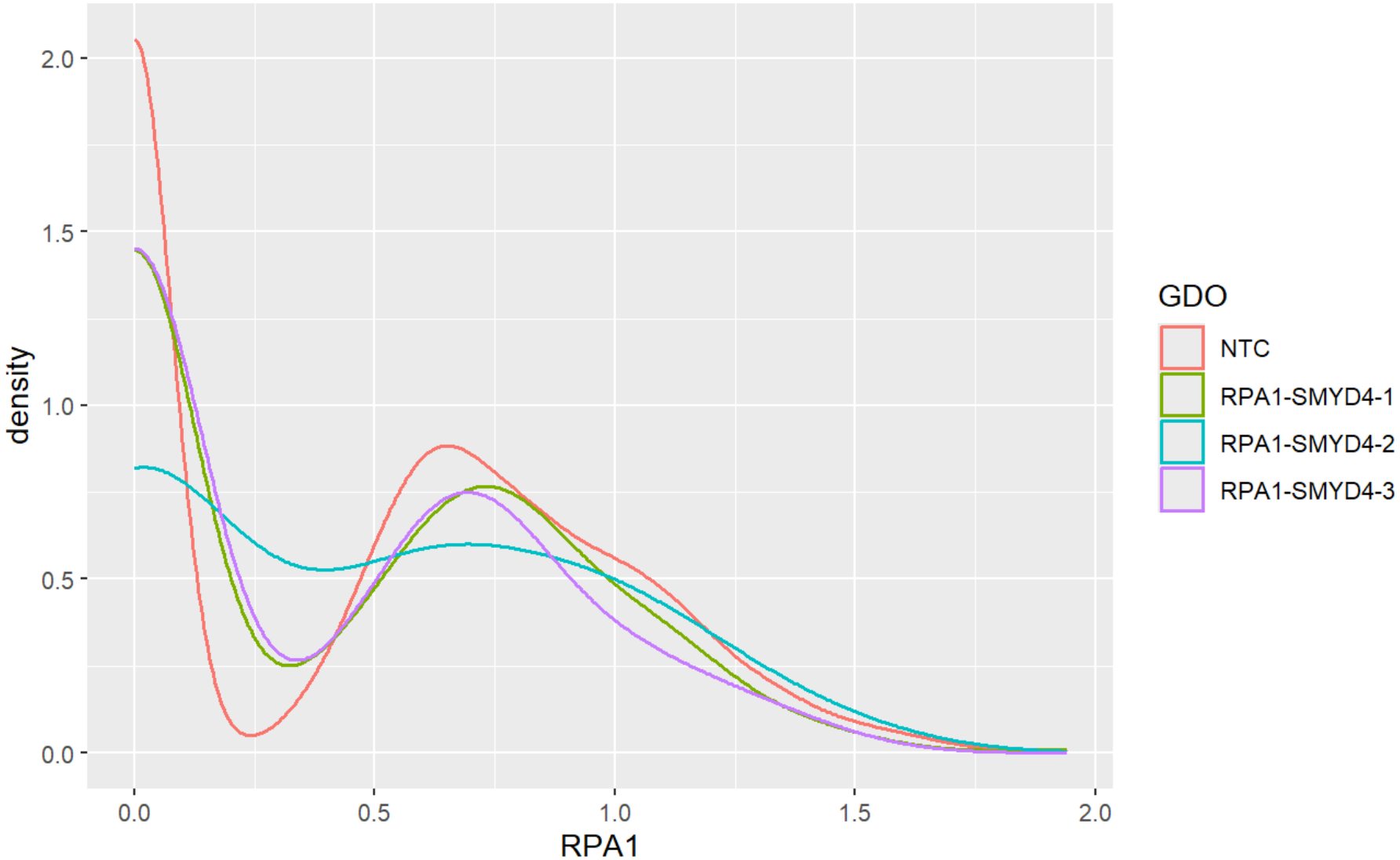
Positive Controls: TSPAN14



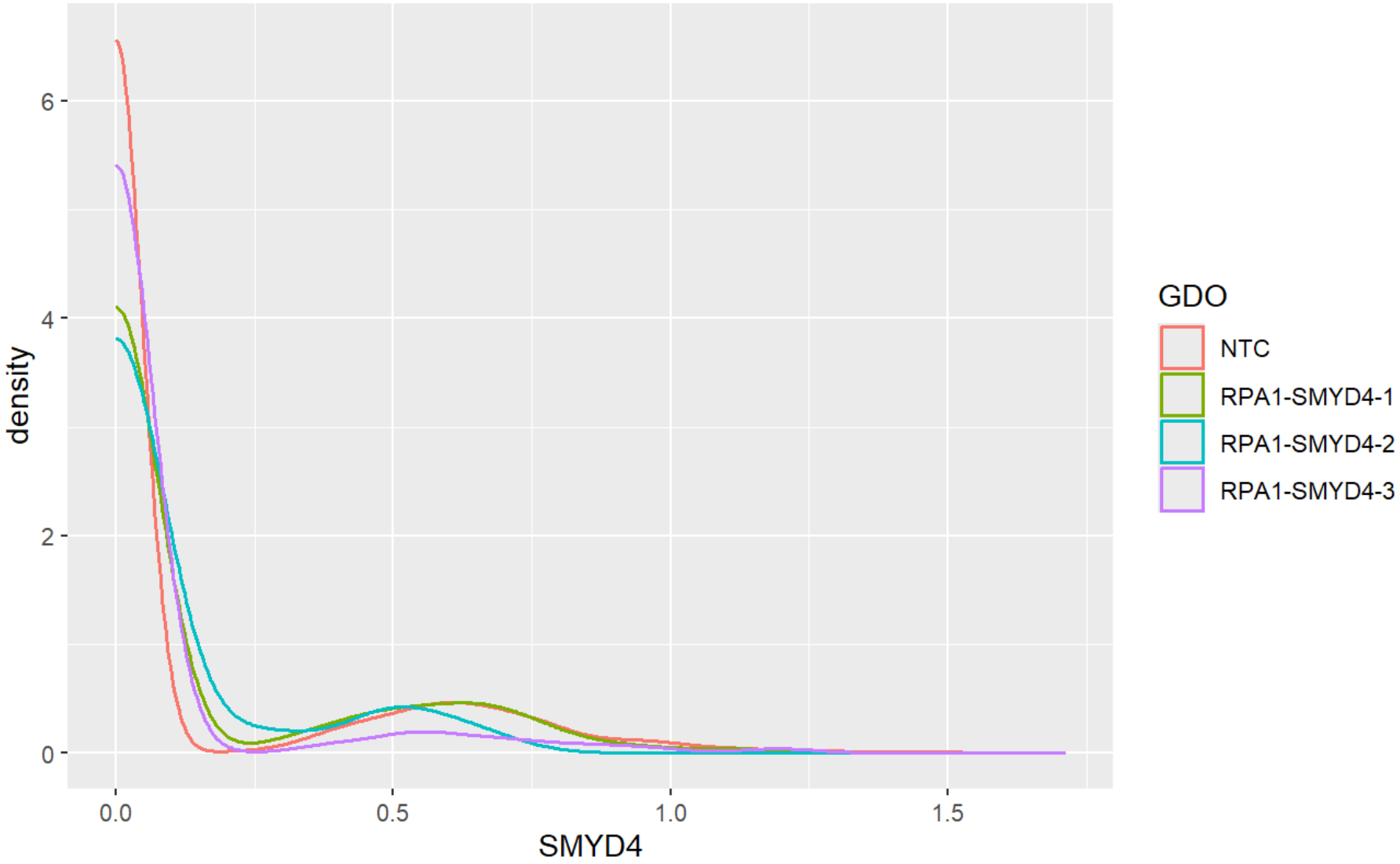
Positive Controls: CLU



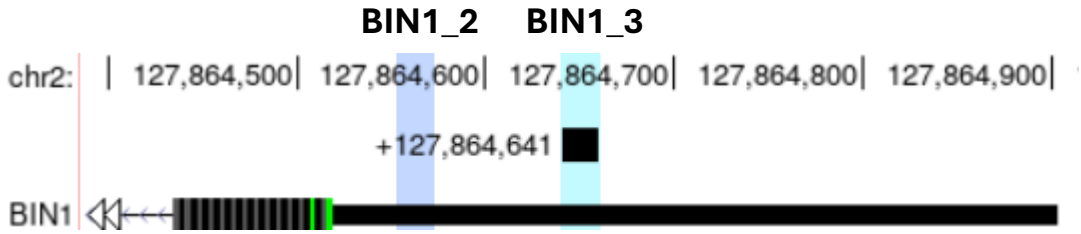
Positive Controls: RPA1

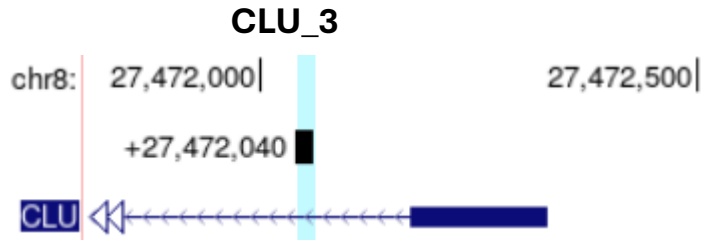


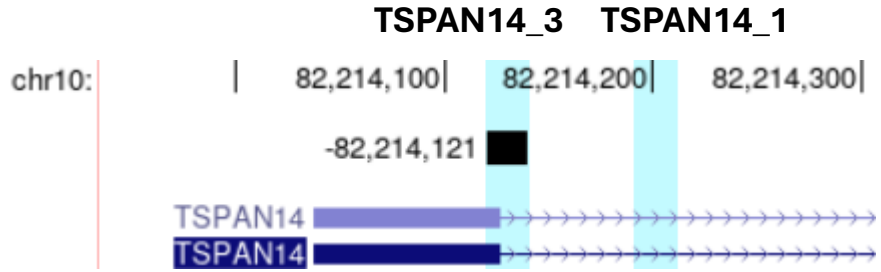
Positive Controls: SMYD4

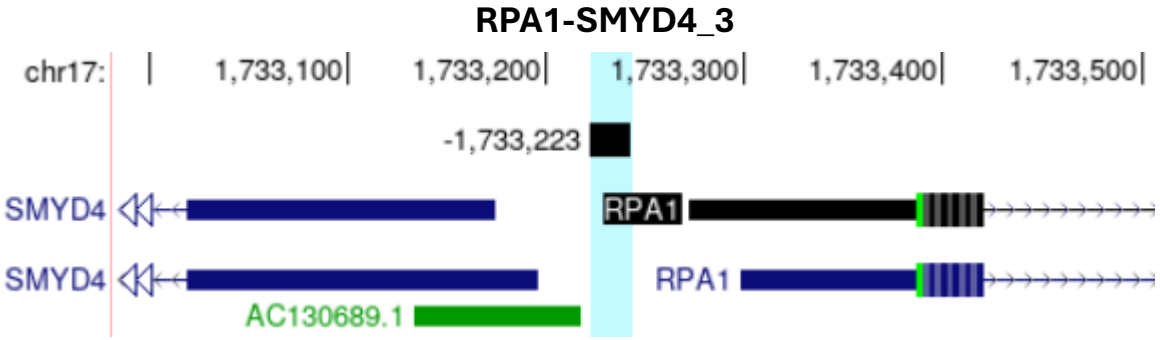


Existing (+) controls & their locations





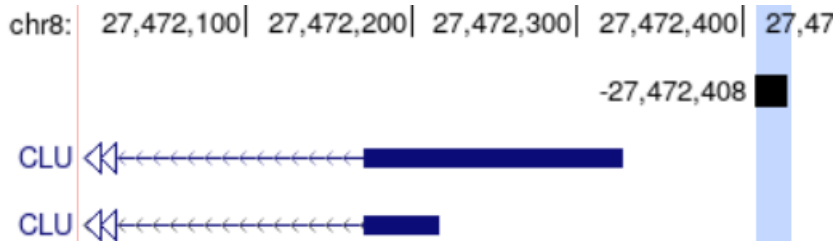




Re: ChatGPT says the skinny, solid box is the 5’ UTR. Still trying to find a source from UCSC Genome Browser.

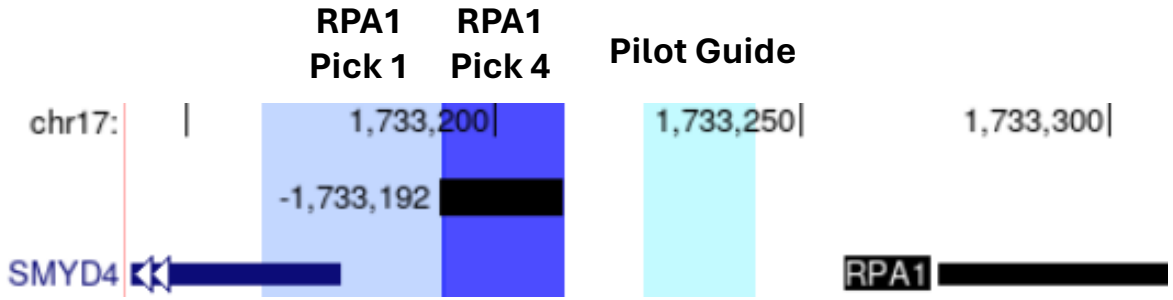
The green line/codon is the Start Codon (confirmed).

Special Cases: CLU guide (non-pilot) and RPA1-SMYD4 guides



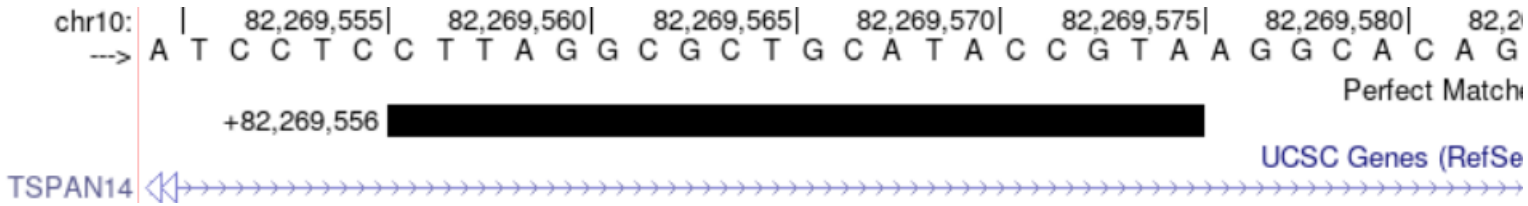
CLU guide 2 from CHOPCHOP that Natalia used in the VP64 and validated via qPCR.

This is NOT a guide from the pilot screen.



The display seems a little strange?
All the guides are the same length..
And Pick 1 and Pick 4 shouldn't technically be overlapping.

Template vs Non-Template



rs708…-4. Aka T3 of the viruses.
Guide Sequence:
CTTAGGCGCTGCATACCGTA
Binds to the non-template strand.

TSPAN14 is on the (+) strand.

Since the guides is the same sequence as the gene, it binds to the opposite strand.

The template strand is used as the template for transcription. The mRNA sequence is the complement of the template.

v Template/+ (TSPAN14 is on (+) strand)

GCTGCAATCCTCcttaggcgctgcataccgtaAGGCACAGCTTCTTC

v Guide is same as (+) strand, **binds** to non-template

sgRNA-4 (screen)

CGACGTTAGGAGgaatccgcgacgtatggcatTCCGTGTCGAAGAAG

^ Non-Template/-